Wei Dai
Fairfax, VA

Email: dwei90bd@gmail.com
Mobile: (216) 762 6960
LinkedIn: wdai144
GitHub: wdai0

SUMMARY

PhD candidate in Statistics (exp. Aug 2025) with a research experience on feature selection, mixture models and stochastic simulation optimization; Experienced R and Python package developer with demonstrated expertise in analyzing electronic health records (EHR) data, applying machine learning methods, and engaging in statistical collaboration. Actively following advancements in large language models.

EDUCATION

PhD in Statistics   **George Mason University, 2020-2025 (exp. Aug)**
- Research Assistant on NSF and NIH funded projects; Teaching Assistant for Statistical Graphics and Data Visualization featuring **R Shiny**, Time Series Analysis, and Regression Models; **Best TA Awardee 2024**.
- Course training in advanced biostatistics, with an emphasis on **clinical trials** and **survival analysis**.

MS in Biostatistics   **Case Western Reserve University, 2018-2020**
- Earned **SAS** certification; maintained 3.90 GPA in statistics and programming coursework.
- Served on **statistical consulting team** providing statistical support to medical school researchers. Demonstrated experience in communication with collaborators with varying levels of statistical knowledge.

PROFESSIONAL EXPERIENCE

GMU Dept. of Statistics & Inova Heart and Vascular Institute   2022-2024
Biostatistician / Research Assistant
- *Developed **predictive models** analyzing OPTN data to forecast one-year heart transplant survival and performed simulation to optimize donor-recipient matching protocols.*
- Participated end-to-end data **pipeline** development, processing sensitive healthcare data on HPC.
- Engineered data pre-processing pipeline, including missing data treatment, and a pilot **survival analysis**, reducing processing time while preserving data integrity.
- Implemented **machine learning models** (logistic regression, XGBoost, neural networks) on **OPTN** data, achieving performance metrics comparable to benchmark publications (on different cohorts).
- Briefed on progress through meetings and **presentations** to team members from diverse backgrounds, conveying analytical insights, culminating in accepted abstract and presentation at **ISHLT2023 conference**.
- paper preprint[2]: *Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant.*

GMU Dept. of Statistics & Cleveland Clinic Critical Care Transport   Jan-June 2021
Data Scientist / Research Assistant
- *EHR diagnosis code subcategorization schema developed to improve **machine learning model applications in clinical settings**. Results provide additional clinical insights and temporal information, improving predictive model performance for clinical decision support applications.*
- Identified and validated six diagnosis subcategories; Normalized **ICD9/ICD10** codes via UMLS identifiers and integrated them with supporting EHR data.
- Implemented **random forest** modeling to evaluate subcategorized versus standard diagnosis codes for mortality prediction at transport. Demonstrated **22% improvement** in testing AUC using subcategorized model versus standard model.
- Published [3] and [1] in *International Journal of Medical Informatics* and in *Front. Disaster Emerg. Med*

GMU Dept. of Statistics & INOVA Division of Cardiology   Jan-Mar 2023
Biostatistician / Research Assistant
- *A statistical analysis of a **longitudinal** Atrial Fibrillation (AFib) recurrence study, combining demographic and biomarker data, revealed two new biomarker associations. Additionallay, the study analyzed factors impacting patient quality of life (AFEQT scores).*
- Focused on explaining factors of AFib recurrence, with a secondary task of analyzing AFEQT (quality of life) scores.
- Progressed beyond t-tests to more sophisticated models including **linear mixed effects model** and **beta regression**.
- Presented work to medical doctors and authored in statistical methods section. paper preprint *Novel Biomarkers to Predict Recurrence of Atrial Fibrillation after Catheter Ablation.*

Doctoral Research

○ Research focuses on variable selection algorithms, mixture models and stochastic simulation optimization.
○ Built **visualization tools** illustrating statistical concepts, improving understanding.
○ Implemented **PyTorch based parallelization** for statistical computing on variable selection, achieving a performance improvement of **20x** over traditional approach.
○ Presented work and poster at JSM 2023 and 2024.
○ Created and published open-source packages `subsampwinner` [6] in both R and Python, available on PyPi and **GitHub**.

SKILLS & INTERESTS

○ Python, R, SAS, SQL, Shell
○ R shiny, PyTorch, Scikit-learn
○ HPC, Git, GitHub, Spark

○ Survival Analysis, Feature Selection, Clinical Trials, Simulation Optimization
○ Machine Learning, Generative AI, Large Language Models

RESEARCH PUBLICATIONS

Articles

[1] Andrew P. Reimer, Wei Dai, Nicholas K. Schiltz, Jiayang Sun, and Siran M. Koroukian. "**Patient Factors Associated with Survival after Critical Care Interhospital Transfer**". In: *Front. Disaster Emerg. Med.* 1 (Jan. 8, 2024), p. 1339798. ISSN: 2813-7302. DOI: 10.3389/femer.2023.1339798.

[2] Jie Xu, Wei Dai, J. Goldberg, P. Shah, I. Hu, C. Chen, C.R. deFilippi, and J. Sun. "**Explainable Machine Learning to Improve Donor-Recipient Matching at Time of Heart Transplant**". In: *The Journal of Heart and Lung Transplantation* 42.4, Supplement (2023). ISHLT 43rd Annual Meeting and Scientific Sessions, S22. ISSN: 1053-2498. DOI: https://doi.org/10.1016/j.healun.2023.02.043.

[3] Andrew P. Reimer, Wei Dai, Benjamin Smith, Nicholas K. Schiltz, Jiayang Sun, and Siran M. Koroukian. "**Subcategorizing EHR Diagnosis Codes to Improve Clinical Application of Machine Learning Models**". In: *International Journal of Medical Informatics* 156 (Dec. 2021), p. 104588. ISSN: 13865056. DOI: 10.1016/j.ijmedinf.2021.104588.

[4] Kath M Bogie, Katelyn Schwartz, Youjin Li, Shengxuan Wang, Wei Dai, and Jiayang Sun. "**Exploring Adipogenic and Myogenic Circulatory Biomarkers of Recurrent Pressure Injury Risk for Persons with Spinal Cord Injury**". In: *J Circ Biomark* 9.1 (Sept. 21, 2020), pp. 1–7. ISSN: 1849-4544, 1849-4544. DOI: 10.33393/jcb.2020.2121.

[5] Dennis Bourbeau, Abby Bolon, Graham Creasey, Wei Dai, Bill Fertig, Jennifer French, Tara Jeji, Anita Kaiser, Roman Kouznetsov, Alexander Rabchevsky, Bruno Gallo Santacruz, Jiayang Sun, Karl B. Thor, Tracey Wheeler, and Jane Wierbicky. "**Needs, Priorities, and Attitudes of Individuals with Spinal Cord Injury toward Nerve Stimulation Devices for Bladder and Bowel Function: A Survey**". In: *Spinal Cord* 58.11 (Nov. 2020), pp. 1216–1226. ISSN: 1362-4393, 1476-5624. DOI: 10.1038/s41393-020-00545-w.

Packages

[6] Wei Dai and Jiayang Sun. *subsampwinner: A package for feature selection using Subsampling Winner Algorithm.* Python Package Index (PyPI). Version 0.0.8, Accessed: 2025-03-08. Aug. 2024.